Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Hidden Pattern of Variation: Mapping SNP Landscapes across Mammalian Genes

Magdalena Fraszczak
Wroclaw University of Environmental and Life Sciences
Paula Dobosz
Poznan University of Medical Sciences
Barbara Karbowa
Wroclaw University of Environmental and Life Sciences
Jakub Liu
Charité - University Medicine Berlin
Magda Mielczarek
Wroclaw University of Environmental and Life Sciences
Weronika Stasiak
Wroclaw University of Environmental and Life Sciences
Joanna Szyda
joanna.szyda@charite.de

Wrocław University of Science and Technology

Research Article

Keywords: gene architecture, mammalian genome, SNP distribution, genomic variation, comparative genomics

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9004337/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: Single nucleotide polymorphisms (SNPs) represent the most abundant form of genetic variation in mammalian genomes and serve as critical markers in evolutionary, functional, and clinical genomics. Despite their extensive use, the distribution of SNPs across genic regions, particularly between exons and introns, remains uneven and not fully characterized across species.

Results: Here, we explored the counts of SNPs in consecutive exons and introns of the human, bovine, and swine genomes, analysing 18,448 human, 19,657 bovine, and 17,342 porcine protein-coding genes (encompassing 41.8 million, 69.2 million, and 23.9 million polymorphisms, respectively). The three species demonstrated a consistent, non-random pattern - excess of SNPs in the first and the last exon as well as the excess of SNPs in the first few introns, especially the 1st one.

Conclusions: The distribution of single nucleotide polymorphisms among introns and exons appears to be not only highly nonuniform but also exhibits a very consistent pattern across mammalian genomes. This observation reflects the distinct functional roles of consecutive exons and introns within a gene.

BACKGROUND

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation. However, their genomic distribution is not random (1,2). Although they are located in all functional genomic elements (promoters, exons, introns, 5'UTRs, 3'UTRs, and intergenic regions), their density varies between regions, with exons and splice sites (marking exon–intron boundaries) being the most conserved, that is, SNP-sparse (3). However, even within functional genomic units, such as introns and exons, the density of SNPs is highly nonuniform (4), with clusters of adjacent SNPs being an often-observed characteristic of the human genome. In particular, Hodgkinson and Eyre-Walker (5) and Prendergast et al. (6) estimated an excess of intronic SNPs located in a single-bp proximity to each other. Matsushita and Kano-Sueoka (7) recently reported differential clustering of synonymous and nonsynonymous SNPs among consecutive exons of the human *HLA-A* gene, while in the whole-genome level, Back and Walther (8) reported a greater density of SNPs in the first intron than in the next introns in *Arabidopsis thaliana*. It has been widely agreed in the literature that such a non-random distribution of SNPs must have an evolutionary impact and is the result of mutational hotspots or that SNP clusters arise due to structural properties of DNA that mechanically promote the accumulation of such point mutations. It is important to note that SNP density and SNP count shall be regarded as two non-equivalent measures of SNP genomic distribution. The SNP count expresses a raw number of polymorphisms identified within a functional unit regardless of the unit length and distances between SNPs, whereas the SNP density refers to the number of SNPs per a given bp interval, hence, both measures are not statistically and biologically equivalent. In our analysis, we focused on SNP count rather than SNP density despite variation in exon and intron length, as the relationship between gene length and SNP density is not always straightforward and can be influenced by gene function, selective pressures, and genomic context (3). Therefore, exons and introns were regarded as functional genomic units rather than simple

stretches of DNA. We explored the number of SNPs in the human, bovine, and porcine genomes, focusing on the differences in SNP counts among consecutive introns and exons. The underlying hypothesis was that there are differences in the numbers of SNPs among particular exons and introns that may reflect their differential role in the formation of the final product of gene expression. For this purpose, we used whole-genome sequence data from humans, represented by 1,222 people; cattle, represented by 5,116 bulls; and swine, represented by 12 pigs.

RESULTS

Genomic distribution of SNPs

Humans, cattle, and swine represent three functionally diverse mammalian species with well-annotated genomes (Table 1).

Species	Number of individuals	Total number of SNPs	Number of analysed SNPs
Homo sapiens	1,222	43,469,928	12,242,273
Bos taurus	5,116	69,222,007	22,172,550
Sus scrofa	12	23,872,646	7,018,226

Table 1. Summary of the analysed whole genome sequence data sets.

humans

Among the 43,469,928 SNPs identified in 1,222 persons 12,242,273 were located in protein coding genes (exons - 700,069, introns – 11,542,204), which made up 28.16% of all the SNPs. Most of the SNPs in both exons (73,250) and introns (1,005,355) were located on HSA01. 18,448 protein-coding genes contained at least one SNP. Most SNPs (54,316) were identified for *CSMD1* (ENSG00000183117) on HSA08, which contains 70 exons. The average number of SNPs per gene was 664±1,459.

cattle

In this dataset, 5,116 bulls were characterized by 69,222,007 SNPs, of which 22,172,550 (32.03%) were located in protein-coding genes (exons – 1,411,623, introns – 20,760,927). Most of the SNPs in both exons (262,674) and introns (1,411,623) were located on BTA02. 19,657 protein-coding genes contained at least one SNP. The highest number of SNPs (63,534) were annotated to *CNTNAP2* (ENSBTAG00000052473) which is located on BTA04 and is composed of 24 exons. On average, there were 1,128±2,705 SNPs per gene.

swine

Among the 23,872,646 SNPs identified in twelve pigs, 7,018,226 (29.39%) were located in protein-coding genes, including 283,004 SNPs in exons and 6,735,222 SNPs in introns. The highest number of exon–

located SNPs (25,691) was found on SSA02, while most of the SNPs located in introns (602,671) were identified on SSA01. 17,342 genes contained at least one SNP. The highest number of SNPs per gene (22,580) was identified within *DLG2* (ENSSSCG00000014904) on SSA09, which is composed of 27 exons. The average number of SNPs per gene amounted to 404±959.

The majority of the analysed genes, that is, 1,518 for swine and 2,535 for cattle, contained exactly two exons, whereas for humans, most of the genes, totalling 1,577, contained four exons (Figure 1).

As expected, in all the considered datasets, exons contained fewer SNPs than introns. However, the percentage of genes with no SNPs in exons was low: 0.06% for cattle, 0.09% for humans, and 5% for swine. In comparison, for introns, the corresponding percentages were 0.1% for humans, 0.4% for pigs and 3% for cattle. The average number of SNPs per exon was the highest in genes with a low number of exons, while the average number of SNPs per intron considerably varied and did not linearly depend on the number of introns (Figure 2). Despite this, in all the considered species the highest average number of SNPs per intron was observed in *HS6ST3* gene (ENSG00000185352, ENSBTAG00000039065, and ENSSSCG00000009503), with 7,495 SNP in the human gene, 22,652 in the cattle gene, and 8,597 in the pig gene. In downstream analyses, only genes containing fewer than 26 exons were considered to preserve class counts for reliable statistical inference.

Differences in the number of SNPs located in exons

The number of SNPs located in exons was significantly nonuniform in all gene groups (i.e., sets of genes with the same number of exons). In addition, in all three species, the differences in SNP counts between exons were significant for all genes with at least three exons (Supplementary Table 1). Pairwise differences in SNP counts between exons were visualized in Figure 3. A very consistent pattern emerged, showing a significant excess of SNPs in the first and the last exons, but the first exon always contained fewer SNPs than did the last one. Similarly, for genes with only two exons, regardless of the species, the second exon contained a higher number of SNPs than the first exon (P-values varied between $2.16\ 10^{-185}$ in humans and $1.37\ 10^{-66}$ in pigs).

Differences in the number of SNPs located in introns

The number of intronic SNPs differed significantly across all gene groups (Supplementary Table 1). Pairwise comparisons of genes with at least 10 introns revealed a consistent pattern in humans, cattle, and pigs, where the first four to five introns contained significantly more SNPs, with the first intron always having the highest count. The second intron had more SNPs than the third, which in turn had more than the remaining introns. Similarly, the fourth and fifth introns contained more variants than the remaining introns of the gene. In gene groups with fewer than 10 introns, a similar tendency was observed; however, in this case, the first three introns exhibited a higher number of SNPs (Figure 4). Furthermore, a higher number of SNPs was found in the first intron compared to the second intron in genes with two introns ($P=2.2\ 10^{-6}$) in humans, and ($P=2.8\ 10^{-4}$) cattle

SNP-rich exons in the porcine genome

While considering the cumulative SNP count in introns and exons, five porcine genes attracted special attention. Despite the lowest number of SNPs identified among the twelve individuals, the genes harboured a very large number of exonic SNPs and no SNPs in introns. The highest number of SNPs (95) distributed across three exons was identified for the novel gene ENSSSCG00000050559 on SSA04. 95 SNPs in three exons were found for a novel gene ENSSSCG00000046109. LOC100157704 (ENSSSCG00000032127), with 91 SNPs in three exons, is related to a G protein-coupled receptor involved in olfactory signalling, contributing to the sensory perception of smell and the detection of chemical stimuli. It plays a central role in GPCR-mediated signal transduction. A novel gene ENSSSCG00000045040 with 56 SNPs in three exons is homologous to the human TMEM258 (ENSG00000134825), which is associated with N-linked glycosylation and participates in protein modification via the oligosaccharyltransferase I complex, essential for proper protein folding and function.

DISCUSSION

Today, very large datasets of SNPs identified from whole-genome sequencing are available, such as the resource provided by the UK Biobank (9). Nevertheless, the human and porcine datasets analysed in our study possess characteristics that make them advantageous for the analysis of SNP genomic distribution. In particular, for both species, individuals were selected and processed as a single cohort and therefore underwent identical methodology of variant calling, including the genotyping platform and sequence pre-processing, which allowed for minimising the technical bias of SNP calling. Moreover, the human dataset represents a timely and geographically uniform group of individuals of Polish origin (10). The porcine dataset consists of individuals housed in one closed piggery, with the same standard environmental, microclimatic, and nutritional conditions. Therefore, this excludes the ascertainment bias of SNP frequency due to population stratification and selection. The bovine data set, albeit not sequenced as a single cohort, underwent a unified SNP calling protocol and is one of the largest bovine data sets currently available that consist of truly sequenced (i.e., non-imputed) whole genome polymorphisms.

SNP distribution

In our study, we deliberately focused on SNP count instead of SNP density, even though the lengths of exons and introns vary considerably. However, while some studies have suggested a correlation between gene length or exon/intron count and SNP density (11), the relationship is not always straightforward. Gene function, selection pressures, and the genomic context can influence SNP counts (3) regardless of their physical proximity. Therefore, in our study, introns and exons were regarded as functional genomic units and not as linear sequences of nucleotides.

The functional role of the genomic region strongly determines the localisation of polymorphisms, since SNPs in exons can have a potential impact on gene products that, on a larger scale, may cause a disease

or alter quantitative phenotypes (12). However, since introns play regulatory roles, the presence of polymorphism in introns may indirectly impact gene products or their expression levels (13). Nevertheless, due to the generally more severe potential consequences of polymorphisms in exons, the expectation is that exons contain fewer SNPs than introns (14), which was confirmed by our study. Moreover, genes with a low number of exons had the highest mean number of SNPs per exon, which was also observed in this study. This may be related to the fact that smaller genes are frequently expressed during an individual's lifetime because they are typically involved in functions that require fast responses, such as the immune system. These specific functions contribute to higher variation, which facilitates the response to and interaction with the dynamic environment (11).

SNP counts in exons

In exonic regions, SNPs can alter protein function directly or affect splicing. Synonymous SNPs, which do not change the amino acid sequence, can still disrupt splicing enhancers or silencers, leading to aberrant splicing patterns (15). For instance, a study revealed that synonymous SNPs are associated with splicing misregulation in diseases (15). In the aforementioned study by Back and Walther (8), which used *Arabidopsis thaliana* as the model genome, a high positive correlation was estimated between sequence variation in the first exons and gene expression. This correlation could explain the higher genomic variability of the first exon observed in our study. Additionally, in a context of across-species comparison based on the reference genome sequence, Castle (4) reported more variability in coding regions in the proximity of the start and stop codons, which typically map to the first and last exons. This finding is in line with our observation of SNP excess in the first and last exons. Moreover, the first and last exons include not only the protein coding sequence but also the 5'- and 3'-untranslated regions (UTRs). UTRs contain a translation initiation codon (5'UTR) and regulatory sequences, including sites for the binding of microRNAs and RNA-binding proteins. These features make them important for RNA stability and mRNA translation, so DNA variation within this region may impact gene expression (16, 17). Consequently, the higher number of SNPs in the first and last exons may be related to the presence of UTRs and express the biological potential to increase variation in gene expression.

SNP counts in introns

The biological role of introns is manifold. They allow for alternative splicing (18) but additionally influence the stability of mRNA (19), contain noncoding genes (20), and include regulatory elements, especially enhancers, that affect the rate of transcription, known as the phenomenon of intron-mediated enhancement (21, 22). Therefore, intronic SNPs, although located in noncoding regions, can influence gene expression. They may disrupt splicing regulatory elements such as intronic splice enhancers or silencers, leading to exon skipping or the inclusion of intronic sequences in mature mRNAs (23). An example is the SNP within intron 4 of the human growth hormone gene, which affects gene expression levels (24). Furthermore, intronic SNPs can be associated with diseases by affecting splicing regulation. A comprehensive analysis revealed that disease-associated intronic SNPs are more likely to disrupt splicing compared to common SNPs, emphasising their potential role in disease mechanisms

(15,25,26). In our study, a significant excess of SNPs was observed in the first introns. Among all introns, the first one has been recognised as having special features and functions, including, among others, correcting the cytoplasmic localisation of some mRNAs, as well as transcriptional and translational regulation (27). The important role of genetic variation in the first introns can also be anticipated by observing a very high number (over 120 since 2001) of publications reporting associations of SNPs located in the first intron with a variety of phenotypes measured in humans, animals, and plants (based on PubMed access on 26.10.2025).). These specific roles may explain why the first intron sequence of human DNA is considered the longest and most highly dense regulatory chromatin mark (28,29). Considering SNP density, the abovementioned studies identified the first introns as the most conserved regions, however, a highly nonuniform distribution of SNPs along the intronic sequence (30), means that the low density of SNPs in some areas doesn't necessarily correspond to a low overall intronic SNP count. Interestingly, in *Arabidopsis thaliana*, Back and Walther (8) reported that the first introns harbour more SNPs than subsequent introns.

CONCLUSION

The distribution of single nucleotide polymorphisms among consecutive introns and exons is not only highly nonuniform but also exhibits a very consistent pattern, with the first introns, first exons, and last exons harbouring significantly more polymorphisms. The same trend was observed regardless of the species and sample size (i.e., the overall number of called SNPs). This observation reflects the important functional role of those genomic units in gene expression by regulating transcription, splicing, or even translation.

METHODS

Materials

The SNP distribution was assessed for three mammalian genomes. SNP variation in the **human genome** was represented by X SNPs identified in 1,222 individuals. The cohort consisted of unrelated individuals of Polish origin from the 1000-Polish Genomes database. The sample consisted of 697 men and 525 women whose ages varied between 2 and 99 years, with a mean age of 45 years. All the samples were collected between April 2020 and April 2021. Details on subject ascertainment, whole-genome sequencing, and variant calling were described by (10). SNP variation in the **bovine genome** was described by 69,222,007 SNPs identified in the genomes of 5,116 bulls, representing various dairy and beef breeds as well as crossbreeds. The majority of individuals represented the Holstein (1,148), Angus (401), and Norwegian Red (347) breeds. The data represents run9 of the 1000 Bull Genome database. The SNP variation of the **porcine genome** comprised a set of 23,872,646 SNPs identified in the genomes of twelve pigs representing the Polish Large White breed.

The human and bovine data sets were accessed at the Variant Calling Format level, with actual SNP calling performed elsewhere within the framework of each database collection, albeit with a sequence

pre-processing and variant calling pipeline unified across both species. The raw whole-genome sequences of pigs (the Illumina HiSeq2000 platform) were processed in-house. The quality control of the raw data was performed with FastQC (31) and MultiQC (32), then the reads were trimmed using Trimmomatic (33). Filtered reads were aligned to the Sscrofa11.1 reference genome by BWA-MEM (34), and postalignment was performed using the SAMtools package (35). As the last step, GATK (36) was used to call SNPs.

SNP processing

Filtering of the original SNP sets was performed using VCFtools software (37). Variants with a mapping quality score below 20, a minimum depth of coverage under 10, a minimum quality of genotypes below 20, and non-biallelic SNPs were discarded. SNPs located within 3 bp of each other were excluded. The remaining SNPs were genomically annotated to introns or exons of canonical transcripts of coding genes using the Ensembl Variant Effect Predictor tool (38).

Data exploration

The statistical analysis pipeline was set up to follow the hypothesis testing scheme of increasing biological specificity, which was applied separately for exons and introns, as well as separately for each group of genes defined by the same number of exons/introns. All analyses were performed separately for each species (humans, cattle, and swine). At each testing step, the null hypothesis was rejected based on the nominal type I error rate ≤ 0.05

1. The null hypothesis of the total number of SNPs being equal among genes was tested using the χ^2 goodness of fit test.
2. The null hypothesis of the number of SNPs being equal in each exon/intron in a group of genes with the same number of exons/introns was tested using the Friedman test (39).
3. For the groups of genes with significant differences in SNP numbers tested in step 2, the null hypothesis of no differences in SNP numbers between each possible pair of exons/introns was tested via the post hoc Conover test (40).
4. For genes with two exons/introns, the null hypothesis of no differences in SNP numbers against the alternative that the second exon/intron contained more/fewer SNPs than the first one was tested via the Wilcoxon signed rank test.

All analyses were performed with the R package, including the libraries: *PMCMRplus, tidyr, dplyr, ggplot2, gridExtra, plotly, hrbrthemes,* and *lattice* (41).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The human data analysed during the current study are available upon request to Paula Dobosz (pdobosz@ump.edu.pl).

The bovine data set was accessed from the 1000 Bull Genomes Run 9.0. Run 8.0 the data is publicly available at the European Nucleotide Archive under the accession PRJEB42783 (https://www.ebi.ac.uk/ena/browser/view/PRJEB42783.

The swine data set is publicly available at the National Center for Biotechnology Information Bioproject database under the accession PRJNA1172736 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1172736).

Competing interests

The authors declare no competing interests.

Funding

The human dataset collection was partially funded by the Polish National Science Centre grant no. SZPITALE JEDNOIMIENNE/2/2020 and by the Medical Research Agency grant no. 2020/ABM/COVID19/0022. The pig dataset was funded by the Wrocław University of Environmental and Life Sciences (Poland) as part of the Ph.D. research program “MISTRZ”, No N090/0005/21.

Authors' contributions

MF and JS conceived and designed the study. MM designed and performed the bioinformatic pipeline. BK, JL, MF, MM and WS performed the formal analysis. MF interpreted and visualized the results. JL, MF, MM, PD and JS wrote the manuscript, All authors read and approved the final manuscript.

Acknowledgements

The computational power was provided by the Poznan Supercomputing and Networking Centre. The authors would like to thank all the sample donors who participated in the study, as well as the medical personnel of the Central Clinical Hospital of the Ministry of the Interior and Administration in Warsaw, for their active support.

References

1. Amos W. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? Proc R Soc B. 2010 May 7;277(1686):1443–9.

2. Neininger K, Marschall T, Helms V. SNP and indel frequencies at transcription start sites and at canonical and alternative translation initiation sites in the human genome. Kalendar R, editor. PLoS ONE. 2019 Apr 12;14(4):e0214816.
3. Deng N, Zhou H, Fan H, Yuan Y. Single nucleotide polymorphisms and cancer susceptibility. Oncotarget. 2017 Dec 15;8(66):110635–49.
4. Castle JC. SNPs Occur in Regions with Less Genomic Sequence Conservation. Ruvinsky I, editor. PLoS ONE. 2011 Jun 6;6(6):e20660.
5. Hodgkinson A, Eyre-Walker A. Human Triallelic Sites: Evidence for a New Mutational Mechanism? Genetics. 2010 Jan 1;184(1):233–41.
6. Prendergast JGD, Pugh C, Harris SE, Hume DA, Deary IJ, Beveridge A. Linked Mutations at Adjacent Nucleotides Have Shaped Human Population Differentiation and Protein Evolution. Zhang G, editor. Genome Biology and Evolution. 2019 Mar 1;11(3):759–75.
7. Matsushita T, Kano-Sueoka T. Non-random Codon Usage of Synonymous and Non-synonymous Mutations in the Human HLA-A Gene. J Mol Evol. 2023 Apr;91(2):169–91.
8. Back G, Walther D. Identification of cis-regulatory motifs in first introns and the prediction of intron-mediated enhancement of gene expression in Arabidopsis thaliana. BMC Genomics. 2021 Dec;22(1):390.
9. Callaway E. World's biggest set of human genome sequences opens to scientists. Nature. 2023 Dec 7;624(7990):16–7.
10. Kaja E, Lejman A, Sielski D, Sypniewski M, Gambin T, Dawidziuk M, et al. The Thousand Polish Genomes—A Database of Polish Variant Allele Frequencies. IJMS. 2022 Apr 20;23(9):4532.
11. Lopes I, Altab G, Raina P, De Magalhães JP. Gene Size Matters: An Analysis of Gene Length in the Human Genome. Front Genet. 2021 Feb 11;12:559998.
12. Nair V, Sankaranarayanan R, Vasavada AR. Deciphering the association of intronic single nucleotide polymorphisms of crystallin gene family with congenital cataract. Indian Journal of Ophthalmology. 2021 Aug;69(8):2064–70.
13. Mukherjee D, Saha D, Acharya D, Mukherjee A, Chakraborty S, Ghosh TC. The role of introns in the conservation of the metabolic genes of Arabidopsis thaliana. Genomics. 2018 Sep;110(5):310–7.
14. Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. Reduced mutation rate in exons due to differential mismatch repair. Nat Genet. 2017 Dec 1;49(12):1684–92.
15. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015 Jan 9;347(6218):1254806.
16. Steri M, Idda ML, Whalen MB, Orrù V. Genetic variants in mRNA untranslated regions. WIREs RNA. 2018 Jul;9(4):e1474.
17. Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, et al. A large-scale binding and functional map of human RNA-binding proteins. Nature. 2020 Jul 30;583(7818):711–9.

18. Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. Alternative splicing and the evolution of phenotypic novelty. Phil Trans R Soc B. 2017 Feb 5;372(1713):20150474.
19. Gupta SK, Carmi S, Ben-Asher HW, Tkacz ID, Naboishchikov I, Michaeli S. Basal Splicing Factors Regulate the Stability of Mature mRNAs in Trypanosomes. Journal of Biological Chemistry. 2013 Feb;288(7):4991–5006.
20. Chorev M, Carmel L. The Function of Introns. Front Gene [Internet]. 2012 [cited 2025 Oct 16];3. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2012.00055/abstract
21. Clancy M, Hannah LC. Splicing of the Maize *Sh1* First Intron Is Essential for Enhancement of Gene Expression, and a T-Rich Motif Increases Expression without Affecting Splicing. Plant Physiology. 2002 Oct 1;130(2):918–29.
22. David-Assael O, Berezin I, Shoshani-Knaani N, Saul H, Mizrachy-Dagri T, Chen J, et al. AtMHX is an auxin and ABA-regulated transporter whose expression pattern suggests a role in metal homeostasis in tissues with photosynthetic potential. Functional Plant Biol. 2006;33(7):661.
23. Cooper DN. Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes. Hum Genomics. 2010;4(5):284.
24. Millar DS, Horan M, Chuzhanova NA, Cooper DN. Characterisation of a functional intronic polymorphism in the human growth hormone (GHI) gene. Hum Genomics. 2010;4(5):289.
25. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. The American Journal of Human Genetics. 2018 Jan;102(1):11–26.
26. Lalonde E, Ha KCH, Wang Z, Bemmo A, Kleinman CL, Kwan T, et al. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. Genome Res. 2011 Apr;21(4):545–54.
27. Jo BS, Choi SS. Introns: The Functional Benefits of Introns in Genomes. Genomics Inform. 2015;13(4):112.
28. Park SG, Hannenhalli S, Choi SS. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genomics. 2014 Dec;15(1):526.
29. Jo SS, Choi SS. Analysis of the Functional Relevance of Epigenetic Chromatin Marks in the First Intron Associated with Specific Gene Expression Patterns. Hurst L, editor. Genome Biology and Evolution. 2019 Mar 1;11(3):786–97.
30. Majewski J, Ott J. Distribution and Characterization of Regulatory Elements in the Human Genome. Genome Res. 2002 Dec 1;12(12):1827–36.
31. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. 2010.
32. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047–8.

33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.
34. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.
35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.
36. Auwera G van der, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra. First edition. Sebastopol, CA: O'Reilly Media; 2020. 467 p.
37. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.
38. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016 Dec;17(1):122.
39. Friedman M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. Journal of the American Statistical Association. 1937 Dec;32(200):675–701.
40. Conover W, Iman R. Multiple-comparisons procedures. Informal report [Internet]. 1979 Feb [cited 2025 Oct 16] p. LA-7677-MS, 6057803. Report No.: LA-7677-MS, 6057803. Available from: https://www.osti.gov/servlets/purl/6057803/
41. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria; 2016. (R Core Team).

Figures

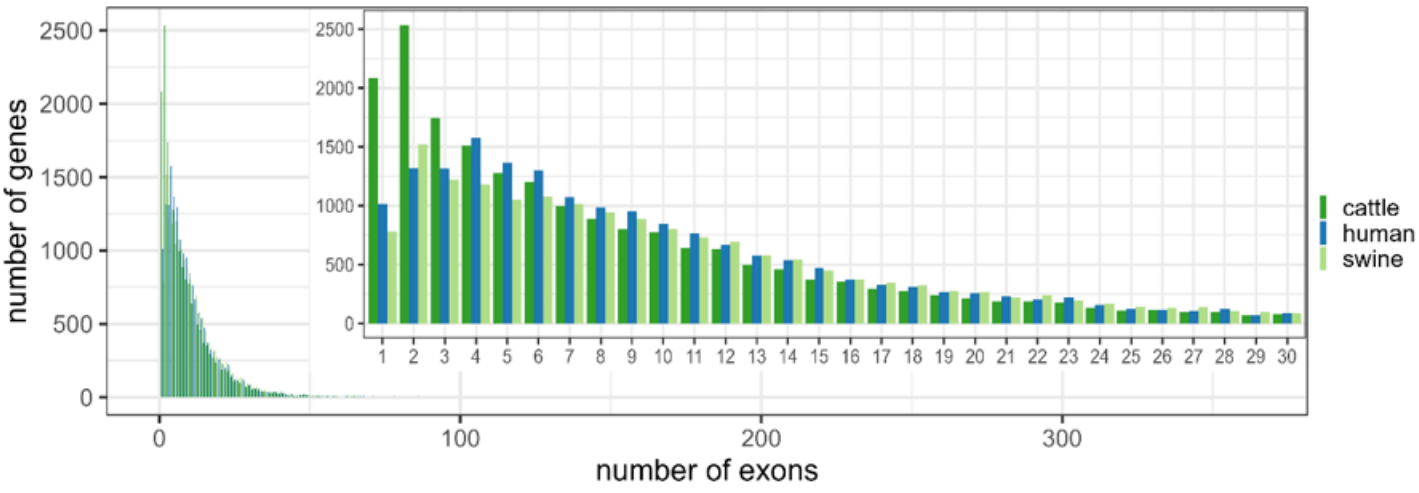


Figure 1

Number of genes depending on number of exons.

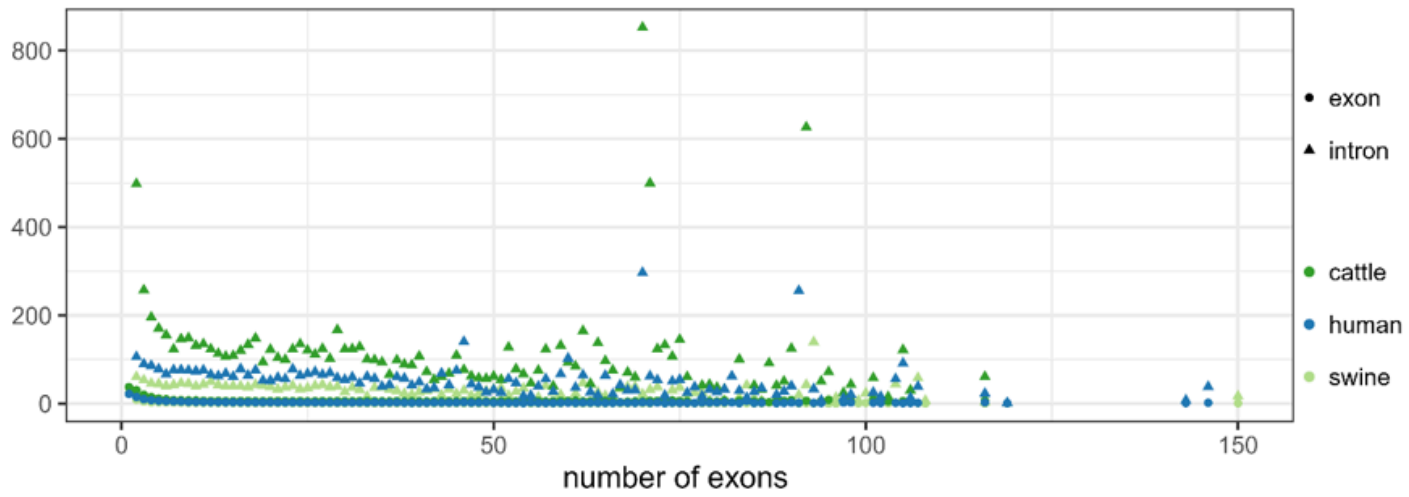


Figure 2

The average number of SNPs located in exons or introns.

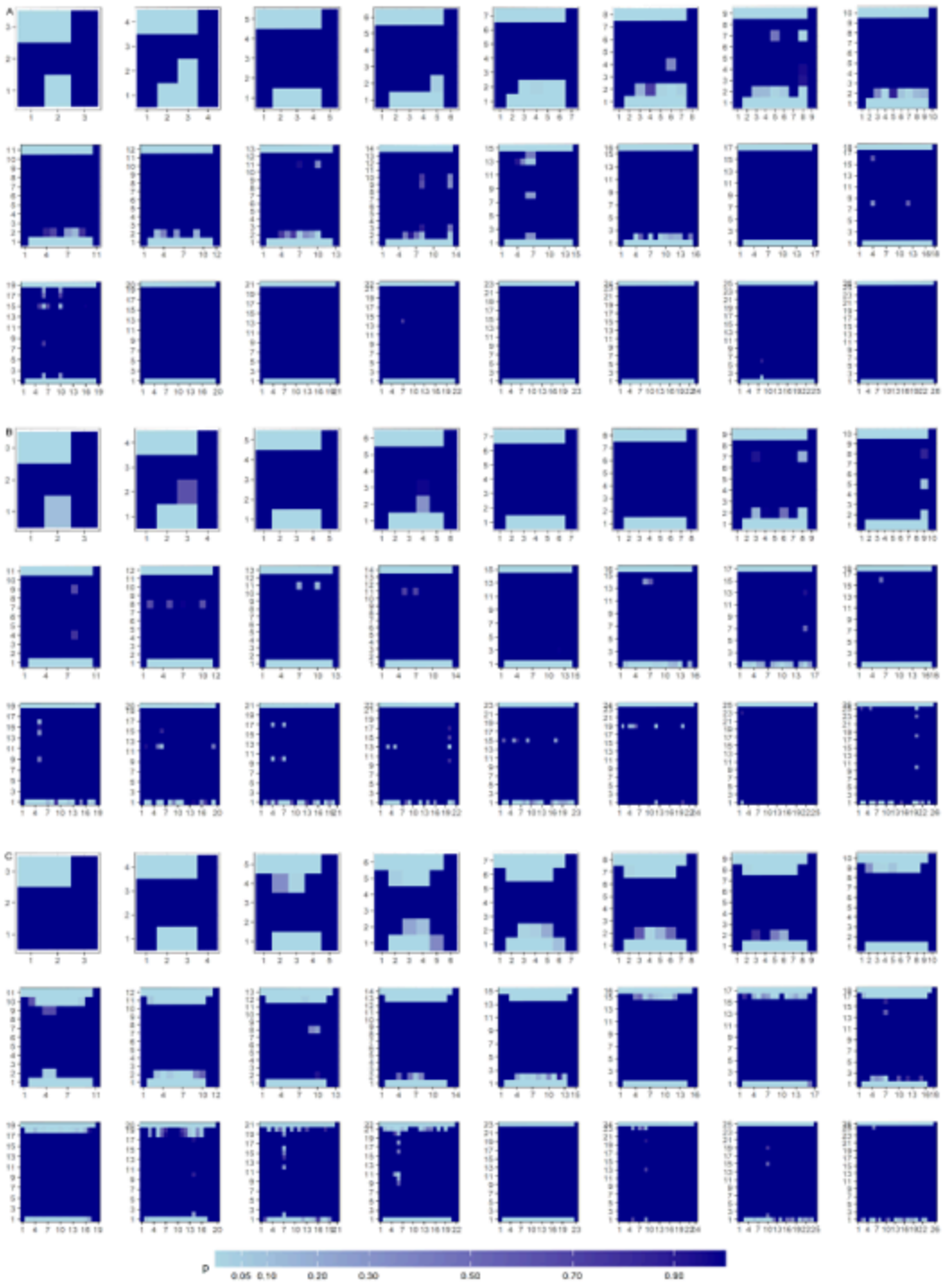


Figure 3

Visualization of the significance of pairwise comparisons of SNP counts across exons. P-values correspond to testing the null hypothesis that the i-th exon (rows, Y-axis) contains more SNPs than the j-th exon (columns, X-axis). (A) humans, (B) bulls, (C) pigs.

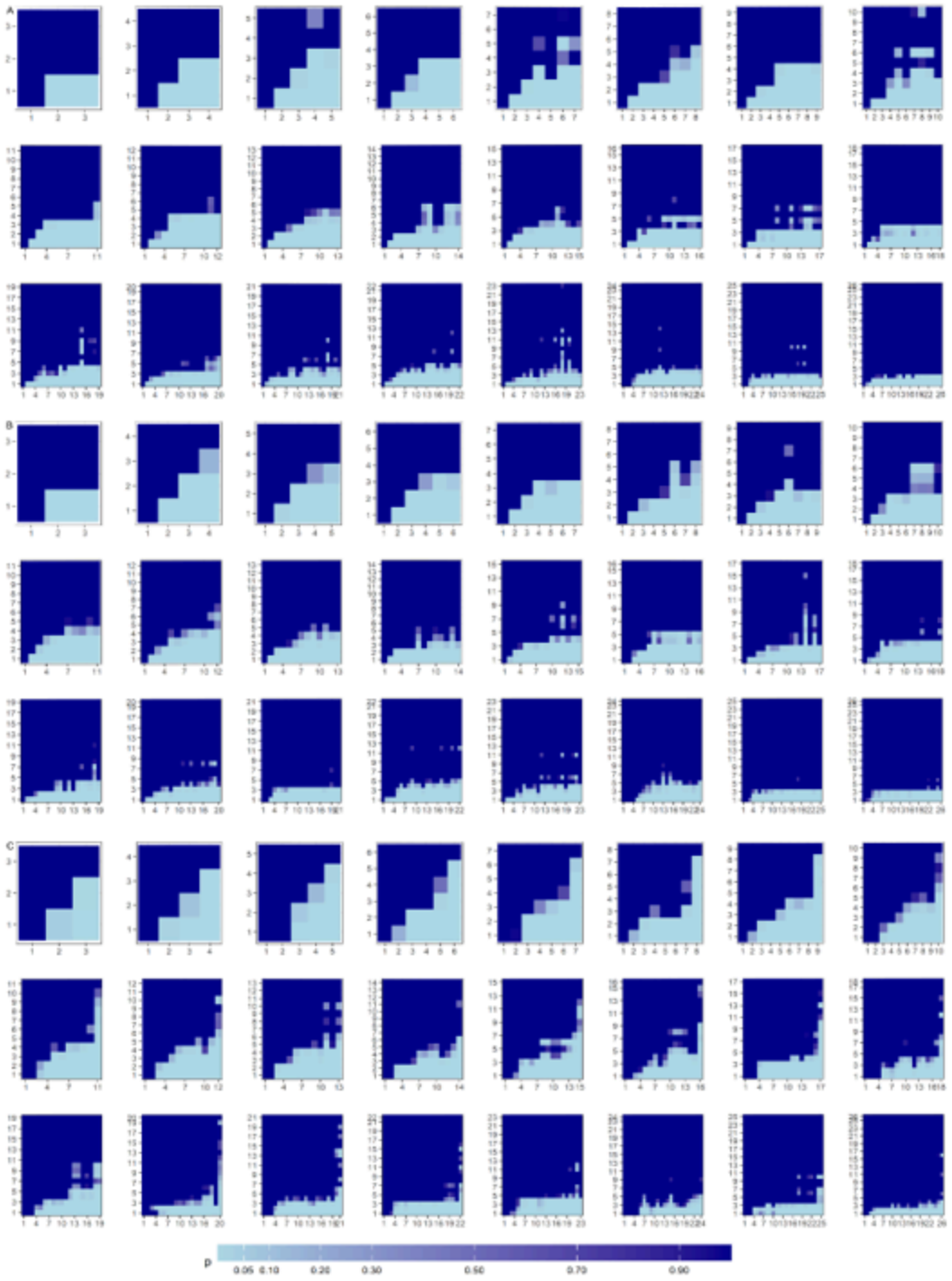


Figure 4

Visualization of the significance of pairwise comparisons of SNP counts across introns. P-values correspond to testing the null hypothesis that the i-th exon (rows, Y-axis) contains more SNPs than the j-th exon (columns, X-axis). (A) humans, (B) bulls, (C) pigs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarytable1.pdf